JP3I (Jurnal Pengukuran Psikologi dan Pendidikan Indonesia)
DOI: http://dx.doi.org/10.15408/jp3i.v13i2.36684
http://journal.uinjkt.ac.id/index.php/jp3i

Comparison of KMO Results, Eigen Value, Reliability, and Standard Error of Measurement: Original & Rescaling Through Summated Rating Scaling

Nuryani Dwi Astuti[12], Mami Hajaroh[1],Yuli Prihatni[2], Ari Setiawan[2], Farida Agus Setiawati[1], Heri Retnawati[1]

[1]Universitas Negeri Yogyakarta, Indonesia
[2]Universitas Sarjanawiyata Tamansiswa Yogyakarta, Indonesia

nuryanidwi.2022@student.uny.ac.id

Abstract

This study aims to compare the results of KMO MSA analysis, Eigen Value, reliability, and Standard Error Measurement (SEm) between raw scores (original) and standardized scores (rescaling) through the summated rating scaling method on critical reasoning attitudes of vocational students in Yogyakarta City. This study used a quantitative descriptive approach involving 204 private vocational students as subjects. The instrument used to measure critical reasoning attitudes has gone through thorough validity and reliability testing before the research was carried out. The analysis process was carried out by calculating the KMO MSA, Eigen Value, reliability, and SEm values on both raw and standardized scores. The results of the two types of scores were then compared to identify any differences. Based on the results of the analysis, it was found that the raw score had a KMO MSA of 0.87, reliability of 0.823, and SEm of 0.337. After rescaling, the KMO MSA value decreased slightly to 0.86, the reliability also decreased slightly to 0.821, while the SEm increased to 0.406. Eigenvalue analysis showed that both the raw and standardized scores yielded seven factors with Eigenvalues greater than 1. The differences found between these two types of scores, namely 0.01 for KMO MSA, 0.002 for reliability, and -0.069 for SEm, indicate small but significant changes, especially in terms of the increase in SEm after rescaling, which impacts the level of measurement accuracy.

Keywords: Reliability, Eigen Value, Summated Rating, Critical Reasoning Attitude.

Abstrak

Penelitian ini bertujuan untuk membandingkan hasil analisis KMO MSA, Eigen Value, reliabilitas, dan Standard Error Measurement (SEm) antara skor mentah (original) dan skor distandarkan (rescaling) melalui metode penskalaan summated rating pada sikap bernalar kritis siswa SMK di Kota Yogyakarta. Penelitian ini menggunakan pendekatan deskriptif kuantitatif dengan melibatkan 204 siswa SMK swasta sebagai subjek. Instrumen yang digunakan untuk mengukur sikap bernalar kritis telah melalui pengujian validitas dan reliabilitas secara menyeluruh sebelum penelitian dilaksanakan. Proses analisis dilakukan dengan menghitung nilai KMO MSA, Eigen Value, reliabilitas, dan SEm baik pada skor mentah maupun distandarkan. Hasil dari kedua jenis skor tersebut kemudian dibandingkan untuk mengidentifikasi perbedaan yang muncul. Berdasarkan hasil analisis, ditemukan bahwa skor mentah memiliki KMO MSA sebesar 0,87, reliabilitas 0,823, dan SEm sebesar 0,337. Setelah rescaling, nilai KMO MSA sedikit menurun menjadi 0,86, reliabilitas juga sedikit menurun menjadi 0,821, sementara SEm mengalami peningkatan menjadi 0,406. Analisis Eigen Value menunjukkan bahwa baik skor mentah maupun distandarkan menghasilkan tujuh faktor dengan nilai Eigen Value lebih besar dari 1. Perbedaan yang ditemukan antara kedua jenis skor ini, yakni 0,01 untuk KMO MSA, 0,002 untuk reliabilitas, dan -0,069 untuk SEm, menunjukkan perubahan kecil tetapi signifikan, terutama dalam hal peningkatan SEm setelah rescaling dilakukan, yang berdampak pada tingkat akurasi pengukuran.

Kata Kunci: *Reliabilitas, Eigen Value, Summated Rating, Sikap Bernalar Kritis.*

Introduction

Critical reasoning is essential for everyday problem-solving, enabling individuals to derive the best conclusions from available facts (Rachmantika & Wardono, 2019). It encompasses the skills to reason, analyze, and apply logic to various scenarios (Abduqodirovich, 2023). However, effective instruction and acquisition of critical reasoning skills in higher education might need to be improved (Effendi et al., 2015; Fajri & Amir, 2022; Golden, 2023). When considering children's reasoning in problem-solving, it is crucial to provide tasks that assess their knowledge, reasoning abilities, and problem-solving strategies (Hamdi et al., 2018; Kim & Pegg, 2019; Pramudita et al., 2019).

Hasanatin & Rohaeti, (2021) created a comprehensive assessment tool to evaluate senior high school students' critical thinking abilities and scientific attitudes, ensuring both high validity and reliability. Cromwell (1992) stressed the necessity of adhering to assessment principles in classrooms to evaluate critical thinking effectively. Crossley (2017, 2016) discussed the difficulties in assessing non-cognitive domains like critical reasoning attitudes due to the challenges in defining these constructs and the inherent subjectivity in assessment. They suggested strategies such as employing cognitive assessments as proxies and objectifying subjective evaluations while also recognizing the limitations of these methods (Allen et al., 1979).

There are four levels of measurement, namely nominal, ordinal, interval, and ratio. Scores resulting from Thurstone and Likert-type measurement instruments are considered ordinal data. Ordinal data analysis only allows the use of the mode and median while not allowing the use of the mean and Standard Deviation (SD). The use of this data has limitations because it does not meet the parametric assumptions related to the data. Kampen (2019) states that the assumptions in parametric analysis involve data obtained at the interval and ratio levels. Therefore, data at the ordinal level cannot be analyzed using parametric statistics, and the analysis carried out must be nonparametric.

The use of ordinal data in parametric data analysis has been a topic of prolonged debate among experts, and until now, an agreement has yet to be reached. Carifio and Perla (2008) noted the debate that has been going on for 50 years regarding data obtained from the Likert measurement model. Jamieson (2004) state that data on a Likert scale has ordinal or ranking characteristics, so it is best to analyze it using nonparametric methods, which are considered less sensitive and more powerful than parametric methods. Suryabrata (2002) noted that data in psychological measurements is not actually interval data but is often treated as interval data. One approach to overcome this debate is through a scaling process, which tries to place attributes or characteristics on a continuum range by changing values or carrying out score transformations, either linearly or nonlinearly (Brennan et al., 2013). In the context of this research, scaling is associated with efforts to change ordinal data, which initially do not have the same comparison units, into interval or ratio data that have equivalent comparison units (Setiawati et al., 2013).

Summated Rating Scaling is important in converting raw scores into standardized scores in educational assessment. This method converts raw data into scores that can be compared in a standardized way, but the results still have ordinal characteristics. In research on multiple intelligence instruments, this method has been shown to be effective in producing score variations in each response (Setiawati, 2013, 2014). Such score variation can affect the variance and standard error of measurement after the data is rescaled (Setiawati, 2013, 2014). It should be noted that in the context of analyses that do not use the Rasch model, the data are not converted to an interval scale but remain in ordinal form. Other methods, such as Rating Scale, have also been used in evaluating community service programs, but retain ordinal characteristics when converting scores through manual input (Wiyono et al., 2017).

In this research, it is important to assess the critical reasoning attitude of vocational students through filling out a questionnaire at the link https://bit.ly/Angket_Bernalar_Kritis.The data generated at this time are still raw scores (original scores before transformation). Therefore, this study uses the summated rating scaling transformation to convert raw scores into standardized scores, which retain ordinal characteristics. The analysis aimed to compare the results of KMO MSA, eigenvalue, reliability, and standard error of measurement between the raw (original) and standardized (rescaled) scores on critical reasoning attitudes of vocational students. However, it should be noted that the rescaled data was not converted into an interval scale because it did not use the Rasch model. The standardized scores used still have ordinal, rather than interval, characteristics.

Methods

The use of descriptive quantitative methods in research is a widely adopted and valuable approach, enabling the summarization and interpretation of data (Clark et al., 2019; Flinton & Malamateniou, 2020). However, there is an observed imbalance in the application of qualitative versus quantitative methods, with the latter being more dominant due to its capacity to measure and generalize findings from larger populations (Vijayendra & Fantone, 2023). Despite this trend, there is an advocacy for a more balanced use of research methodologies, including descriptive statistics, to enhance research across various fields (Proches, 2016). This instrument consists of 36 statement items that have been tested for validity and reliability (Retnawati, 2016). Validity using Aiken was 0.91 in the "very high" category, and reliability using Cronbach's Alpha was 0.823 in the "strong" category.

Instruments were then scaled using a Summated Rating, which looked for KMO MSA, Eigen Value, reliability, and standard error of measurement (SEm). This was done to compare the results of KMO MSA, Eigen Value, reliability, and Standard Error of measurement (SEm) between raw scores (original) and standardized scores (rescaling) through summated rating scaling of attitudes reason critical vocational school students.

The subject study consists of 204 private vocational school students in Yogyakarta City. Data is collected through a spread questionnaire using the link at https://bit.ly/Angket_Bernalar_Kritis. After the data is collected, the step is data analysis.

This deep data analysis study uses statistical calculations with the help of Microsoft Excel, SPSS 26, and R programs. The calculations compare the results of KMO MSA, Eigen Value, reliability, and Standard Error of Measurement (SEm) between raw scores (original) and standardized scores (rescaling) through summated rating scaling of the critical thinking attitudes of vocational school students.

The steps for analyzing the results of KMO MSA, Eigen Value, reliability, and SEm between raw and standardized scores follow the guidelines set by ElHafeez et al. (2022) and Sabon et al. (2022). These steps include 1) Calculate the number frequency (f) of subjects on all responses for each Item. 2) Calculate the proportion (p) by dividing the frequency (f) by the number of respondents (N). 3) Calculate the cumulative proportion (cp), which includes the proportion in each category plus the previous proportion category. 4) Calculate the middle-cp value, which is the midpoint of the cumulative proportion. 5) Calculate the sign deviation (z) using the normal z curve table. 6) Determine the smallest value deviation of 0 by summing the values until it reaches the lowest value of 0. 7) Summing up all value categories plus the smallest value deviation in step 6. Meanwhile, for reliability calculations, use the reliability test using Cronbach's Alpha technique (Christmann & Aelst, 2006; Purnamasari et al., 2020;Sharma, 2016; Taber, 2018). The reliability coefficient formula is as follows:

$$\boldsymbol{\alpha} = \left(\frac{K}{K-1}\right)\left(\frac{S_r^2-\sum S_i^2}{S_x^2}\right) \qquad \text{(Formula 1)}$$

where:

α (alpha) = Cronbach's Alpha Reliability

K = Amount Item tested questions

$\sum$ = Amount variant score Item

S = Variation scores test (entire Item K)

If the alpha value exceeds 0.7, this indicates that reliability is adequate (Faralina et al., 2016; Prasasti & Istiyono, 2018; Setiawan et al., 2024; Sugiyono et al., 2017; Sumin et al., 2022; Syahputra et al., 2023; Wardani et al., 2018). In addition, if alpha reaches more than 0.8, it can be concluded that all test items are considered reliable with a high level of reliability (Amin et al., 2018; Sugiyono et al., 2017). Reliable test instrument standards are determined based on a comparison of the alpha value with the reference value (r table). If the alpha value is greater than the r table, then the question is considered reliable. In contrast, if the alpha value is smaller than the r table, then the question is considered unreliable. In evaluating test instruments, this comparison is the main criterion for determining the reliability of the instrument after the Cronbach's Alpha reliability coefficient is calculated and compared with established reliability standards (Clements et al., 2008; Nolan et al., 2012; Kalkbrenner, 2021; Teng & Wang, 2021).

This initial step is carried out by looking for matrix relationships between the observed indicators. Several metrics can be used as guidelines to ensure data completeness. The application of SPSS software for data analysis has been extensively examined across various fields. Jian-Hua (2005) emphasize the efficiency and accuracy of SPSS in medical and sports measurement studies, respectively. Additionally, MacInnes (2017) illustrates the predictive capabilities and cost-saving benefits of SPSS in base station construction. Stępień et al. (2021) also highlight the reliability of SPSS in statistical analysis, particularly for sphericity measurements. Collectively, these studies highlight the versatility and effectiveness of SPSS in diverse research contexts.

However, in this study, we chose to use the Kaiser-Meyer Olkin (KMO) measurement method to analyze data using SPSS software. This method is generally used to assess the completeness of data required for factor analysis. The KMO method is used to measure the completeness of the sample, as well as to assess the completeness of the sample for each indicator. This method checks the homogeneity of indicators, and the results can be found in Table 3.4, in accordance with recommendations from Kaiser as expressed by previous studies (Alotaibi, 2017; Hudha & Mardapi, 2018; Kalkbrenner, 2021; Kumolohadi et al., 2021; Okoye et al., 2021; Puspitasari et al., 2019; Rasool et al., 2021; Santoso et al., 2022).

In general, a high KMO is very necessary. The higher the KMO value, the better the factor analysis determination (Kurniawan & Munadi, 2019; Rosana et al., 2020). Judging from Table 3.1 above, at least the KMO value can be above 0.80. However, values above 0.50 can usually still be accommodated for determining factor analysis (Hajaroh et al., 2021).

The basic formula for calculating KMO is as follows:

$$KMO = \frac{\Sigma\Sigma r_{ij}^2}{\Sigma\Sigma r_{ij}^2 + \Sigma\Sigma a_{ij}^2} \qquad \text{(Formula 2)}$$

Where: $\mathrm{r}_{\mathrm{ij}}^2$ = koefisien korelasi

$\mathrm{a}_{\mathrm{ij}}^2$ = koefisien korelasi parsial

Apart from including all indicators in the correlation calculation, KMO also calculates the correlation coefficient in the factor analysis for certain indicators; the formula is as follows:

$$MSA = \frac{\Sigma r_{ij}^2}{\Sigma r_{ij}^2 + \Sigma a_{ij}^2} \qquad \text{(Formula 3)}$$

Where: $\mathrm{r}_{\mathrm{ij}}^2$ = koefisien korelasi

$\mathrm{a}_{\mathrm{ij}}^2$ = koefisien korelasi parsial

Kaiser-Meier Olkin's Measure of Sampling Adequacy (KMO) is crucial for assessing the suitability of data for factor analysis (Nkansah, 2018). Generally, a KMO value between 0.6 and 0.7 is deemed acceptable, but the connection between individual KMO values and commonality can be complex (Tyler & Michael, 1958). Kaiser's measure of sampling adequacy (MSA) for single variables is also significant, with a higher MSA correlation coefficient indicating better suitability for factor analysis. Despite its usefulness in identifying unsuitable items before factor analysis, MSA needs to be more noticed (Lorenzo-Seva & Ferrando, 2021). Meyer (1977) emphasizes the importance of a high MSA value, recommending it be at least 0.5 for data to be considered appropriate for factor analysis.

Results and Discussion

The study emphasizes the importance of critical reasoning skills in everyday life and academic settings. By honing these skills, students can make well-informed decisions, solve problems systematically, and draw contextually accurate conclusions (Rachmantika & Wardono, 2019). The assessment of critical reasoning attitudes involves assigning values to non-cognitive attributes, which can reflect the qualifications of the measured attributes (Allen et al., 1979).

The study explains the four levels of measurement: nominal, ordinal, interval, and ratio. Ordinal data, derived from instruments like the Likert scale, traditionally allow only for mode and median calculations and are typically analyzed using nonparametric methods due to their limitations in meeting parametric assumptions (Kampen, 2019). However, the transformation of ordinal data into interval or ratio data through scaling methods has been a topic of extensive debate by Carifio and Perla (Setiawati et al., 2013).

The research utilizes summated rating scales, known for their reliability and validity in measuring attitudes and opinions (Alderson et al., 1992). Various scaling methods, such as stimulus-centered, response-centered, and subject-centered scaling, are employed to convert ordinal data into interval data, which allows for more detailed statistical analysis Torgerson (Setiawati et al., 2013).

The study used an adapted instrument consisting of 36 items, validated and tested for reliability, with a validity score of 0.91 (very high) and a reliability score of 0.823 (strong). The instrument was then scaled using summated rating scaling to compare the results of KMO MSA, Eigen Value, reliability, and Standard Error of Measurement (SEm) raw scores (original) and standardized scores (rescaling) through rating scaling.

Data from 204 vocational school students were collected via a questionnaire and analyzed using Microsoft Excel, SPSS 26, and the R Program. The analysis steps included calculating frequencies, proportions, cumulative proportions, and z-scores for summated rating scaling. Reliability was assessed using Cronbach's Alpha, and the KMO MSA was calculated using SPSS and the R Program.

The study presents calculations for five of the 36 items to illustrate the scaling process. For each Item, frequencies and proportions were calculated, and z-scores were derived to transform ordinal responses into interval data. This transformation resulted in varied response scores, allowing for finer data granularity and potentially affecting descriptive statistics, validity, and reliability.

The comparison of KMO MSA, Eigen Value, reliability, and SEm raw scores (original) and standardized scores (rescaling) through summated rating scaling showed minimal differences, suggesting that the scaling method does not significantly impact the overall results. The KMO MSA values for individual items' raw scores (original) and standardized scores (rescaling) through summated rating scaling remained relatively consistent, indicating that scaling did not markedly improve or degrade the factor analysis suitability.

Tables 1 through 5 demonstrate the summated rating scaling calculations for selected items. The comparison tables (Tables 6-9) show the results of raw scores (original) and standardized scores (rescaling) through summated rating scaling, with a negligible difference in overall KMO MSA values (0.87 original and 0.86 rescaling). This indicates that the summated rating scaling method does not significantly affect the KMO MSA results, suggesting its limited impact on the instrument's psychometric properties.

Calculation Summated Rating Scaling

The initial steps in Summated Rating Scaling, as described by Spector (1992), include calculating the frequency and proportion of responses for each Item, followed by determining the cumulative proportion and median value. This process is essential for creating reliable and valid rating scales, as discussed by (Brown & Daniel, 1990; Shevlin et al., 1997). Shevlin et al.'s (1997) research underscores the effect of reliability on the precision of composite scores, while Brown compares different scaling methods and their underlying assumptions. Spector (1976) stresses the importance of using equal interval response categories to maintain the scale's accuracy. These procedures are critical in developing effective rating scales.

The initial steps taken in Summated Rating Scaling are 1) Calculate the number frequency (f) of subjects on all responses for each Item. 2) Calculate the proportion (p) by dividing the frequency (f) by the number of respondents (N). 3) Calculate the cumulative proportion (cp), which includes the proportion in each category plus the previous proportion category. 4) Calculate the middle-cp value, which is the midpoint of the cumulative proportion. 5) Calculate the sign deviation (z) using the normal z curve table. 6) Determine the smallest value deviation of 0 by summing the values until it reaches the lowest value of 0. 7) Summing up all value categories plus the smallest value deviation in step 6 (Spector, 1992).

The article includes the calculation for summated rating scaling based on 36 items, but only 5 items are shown to save space. The complete calculations for all 36 items can be accessed via this link: https://bit.ly/36_Item_Summated_Rating. Tables 1 through 5 display examples of calculations for the 5 items (specifically items 1, 10, 18, 27, and 36).

Table 1. Calculation of Summated Rating Scaling for Item 1

Item 1	**Response**			
	1	**2**	**3**	**4**
f	32	143	25	4
p	0.15686	0.70098	0.12255	0.01961
cp	0.15686	0.85784	0.98039	1.00000
middle-cp	0.07843	0.50735	0.91912	0.99020
z	-1.41570	0.01843	1.39916	2.33377
z+	1.00000	2.43413	3.81486	4.74947

Sources: Personal data (2022)

In Table 1. the summated rating scaling score is generated from the z score for each response at each point. These results indicate that by scaling, the response score for each point is different from the response score without scaling. The scaled scores show the variation in scores between responses at each point, which are not fixed or always equal to 1. For example, response 1's score changes to -1.41570; response 2 to 0.01843; response 3 to 1.39916; response 4 to 2.33377. If the lowest score is changed to 1.00000, then score 2 becomes 2.43413, score 3 becomes 3.81486, and score 4 becomes 4.74947. The impact of the variation in response scores on the summated rating scale includes finer variations in the data, influence on descriptive statistics (mean, variance, and normality of data), validity and reliability results, and interpretation of research results.

Table 2. Summated Rating Scaling Calculation in Item 10

Item 10	Response			
	1	2	3	4
f	15	75	92	22
p	0.07353	0.36765	0.45098	0.10784
cp	0.07353	0.44118	0.89216	1.00000
middle-cp	0.03676	0.25735	0.66667	0.94608
z	-1.78953	-0.65153	0.43073	1.60796
z+	1.00000	2.13800	3.22026	4.39749

Sources: Personal data (2022)

In Table 2, the summated rating scaling score is generated from the z score for each response at each point. These results indicate that by scaling, the response score for each point is different from the response score without scaling. The scaled scores show the variation in scores between responses at each point, which are not fixed or always equal to 1. For example, response 1's score changes to -1.78953; response 2 to -0.65153; response 3 to 0.43073; response 4 to 1.60796. If the lowest score is changed to 1.00000, then score 2 becomes 2.13800, score 3 becomes 3.22026, and score 4 becomes 4.39749. The impact of response score variation on the summated rating scale includes finer granularity of data, changes to descriptive statistics (such as mean, variance, and normality of data), validity and reliability results, and interpretation of research results.

Table 3. Summated Rating Scaling Calculation in Item 18

Item 18	Response			
	1	2	3	4
f	31	73	90	10
p	0.15196	0.35784	0.44118	0.04902
cp	0.15196	0.50980	0.95098	1.00000
middle-cp	0.07598	0.33088	0.73039	0.97549
z	-1.43264	-0.43748	0.61400	1.96842
z+	1.32722	2.32238	3.37386	4.72828

Sources: Personal data (2022)

In Table 3, the summated rating scaling score is generated from the z score for each response at each point. These results indicate that by scaling, the response scores for each point are different from the response scores without scaling. The scaled scores show the variation in scores between responses at each point, which are not fixed or always equal to 1. For example, response 1's score changes to -1.43264; response 2's to -0.43748; response 3's to 0.61400; response 4's to 1.96842. If the lowest score is changed to 1.32722, then score 2 becomes 2.32238, score 3 becomes 3.37386, and score 4 becomes 4.72828. The impact of response score variation on the summated rating scale includes finer details of the data, changes in descriptive statistics (such as mean, variance, and normality of data), validity and reliability results, and interpretation of research results.

Table 4. Summated Rating Scaling Calculation in Item 27

Item 27	Response			
	1	**2**	**3**	**4**
f	7	65	79	53
p	0.03431	0.31863	0.38725	0.25980
cp	0.03431	0.35294	0.74020	1.00000
middle-cp	0.01716	0.19363	0.54657	0.87010
z	-2.11637	-0.86461	0.11700	1.12685
z+	0.49160	2.57494	3.55654	4.56640

Sources: Personal data (2022)

In Table 4, the summated rating scaling score is generated from the z score for each response at each point. These results indicate that by scaling, the response scores for each point are different from the response scores without scaling. The scaled scores show the variation in scores between responses at each point, which are not fixed or always equal to 1. For example, response 1's score changes to -2.11637; response 2's to -2.11637; response 3's to 0.11700; response 4's to 1.12685. If the lowest score is changed to 0.49160, then score 2 becomes 2.57494, score 3 becomes 3.55654, and score 4 becomes 4.56640. The impact of response score variation on the summated rating scale includes refinement of data details, changes to descriptive statistics (such as mean, variance, and data normality), validity and reliability results, and interpretation of research results.

Table 5. Summated Rating Scaling Calculation in Item 36

Item 36	Response			
	1	**2**	**3**	**4**
f	14	51	99	40
p	0.06863	0.25000	0.48529	0.19608
cp	0.06863	0.31863	0.80392	1.00000
middle-cp	0.03431	0.19363	0.56127	0.90196
z	-1.82086	-0.86461	0.15420	1.29281
z+	0.93900	1.89525	2.91406	4.05267

Sources: Personal data (2022)

In Table 5, the summated rating scaling score is generated from the z score for each response at each point. These results indicate that by scaling, the response score for each point is different from the response score without scaling. The scaled scores show the variation in scores between responses at each point, which are not fixed or always equal to 1. For example, response 1's score changes to -1.82086; response 2's to -0.86461; response 3's to 0.15420; response 4's to 1.29281. If the lowest score is changed to 0.93900, then score 2 becomes 1.89525, score 3 becomes 2.91406, and score 4 becomes 4.05267.

By considering the process of scaling the instrument through the methods described in Table 1 through Table 5 and in full through https://bit.ly/36_Item_Summated_Rating, it can be concluded that the impact of response score variation on the summated rating scale includes finer data granularity, changes in descriptive statistics (such as mean, variance, and data normality), validity and reliability results, and interpretation of research results. In addition, scaling Likert-type instruments with the summated rating method is actually a scaling process with a response approach. In this study, an attempt was made to compare the results of KMO MSA, Eigen Value, reliability, and Standard Error Measurement (SEm) raw scores (original) and standardized scores (rescaling). The instrument was scaled. In classical theory, instruments are analyzed using the summated rating method.

Comparison of Analysis Results Each KMO MSA item, Eigen Value, Coefficient Reliability, and Standard Error of Measurement (SEm)

Before comparing KMO MSA, Aigen Value, coefficient reliability, and Standard Error of Measurement (SEm), KMO MSA analysis using the R Program using Standard Error of Measurement (SEm). Help MS devices. Excel, and coefficient's reliability using SPSS. Step First is to see the respondents' raw scores (original) and standardized scores (rescaling) through summated rating scaling. As for results analysis using SPSS, it can be explained as follows.

Table 6. Analysis Results Raw Scores

Case Processing Summary			
		N	%
Cases	Valid	204	100.0
	Excluded [a]	0	,0
	Total	204	100.0

a. Listwise deletion based on all variables in the procedure.

Sources: Personal data (2022)

Table 7. Analysis Results Standardized Scores

Case Processing Summary			
		N	%
Cases	Valid	204	100.0
	Excluded [a]	0	,0
	Total	204	100.0

a. Listwise deletion based on all variables in the procedure.

Sources: Personal data (2022)

Based on Table 6 and Table 7, it can be explained that 204 students were the respondents from the analysis of raw scores (original) and standardized scores (rescaling) through summated rating scaling, and all respondents' answers were filled in so that the valid number was 100%.

Comparison of the overall KMO (Kaiser-Meyer-Olkin) Measure of Sampling Adequacy (MSA) analysis results between the original raw scores and the scores that have been standardized through rescaling was obtained using the summed rank scaling method with the R program. These results show minimal differences between the two methods, with KMO values of 0.87 for the original scores and 0.86 for the rescaled scores, indicating that both assessment methods have the same high level of adequacy for factor analysis.

Based on the comparison of the overall KMO analysis results in raw scores (original) and standardized scores (rescaling) through summated rating scaling, a difference of 0.01 was obtained. Overall, the scaling method actually reduces the KMO MSA value. This means that the scaling method does not have a significant effect on the KMO MSA results on an instrument, in this case, the critical reasoning attitude instrument for vocational school students.

Furthermore, Table 8 compares the results of the KMO MSA analysis for each Item's raw scores (original) and standardized scores (rescaling) through summated rating scaling using the R program.

Table 8. Comparison of KMO MSA analysis results for each Item original and rescaling

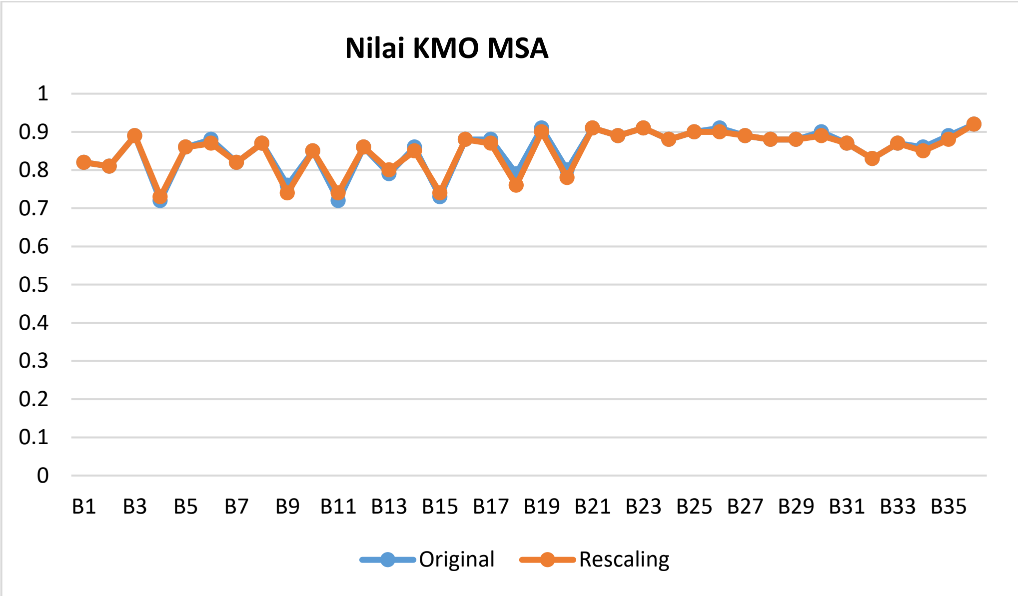
Item	KMO MSA	
	Original	Rescaling
B1	0.82	0.82
B2	0.81	0.81
B3	0.89	0.89
B4	0.72	0.73
B5	0.86	0.86
B6	0.88	0.87
B7	0.82	0.82
B8	0.87	0.87
B9	0.76	0.74

Item	KMO MSA	
	Original	Rescaling
B10	0.85	0.85
B11	0.72	0.74
B12	0.86	0.86
B13	0.79	0.80
B14	0.86	0.85
B15	0.73	0.74
B16	0.88	0.88
B17	0.88	0.87
B18	0.79	0.76
B19	0.91	0.90
B20	0.80	0.78
B21	0.91	0.91
B22	0.89	0.89
B23	0.91	0.91
B24	0.88	0.88
B25	0.90	0.90
B26	0.91	0.90
B27	0.89	0.89
B28	0.88	0.88
B29	0.88	0.88
B30	0.90	0.89
B31	0.87	0.87
B32	0.83	0.83
B33	0.87	0.87
B34	0.86	0.85
B35	0.89	0.88
B36	0.92	0.92

Sources: Personal data (2022)

Based on the analysis using the R Program, the KMO MSA value for each Item raw scores (original) and standardized scores (rescaling) through summated rating scaling increased by four items (B4, B11, B13, and B15), 11 items decreased (B6, B9, B17, B18, B19, B20, B26, B30, B34, and B35), and 21 fixed items (B1, B2, B3, B5, B7, b8, B10, B12, B14, B 16, B21, B22, B23, B24, B25, B27, B28, B29, B31, B32, B33, and B36). So overall, the average KMO MSA value before scaling was 0.87, and after being scaled with a summated rating, it was 0.86.

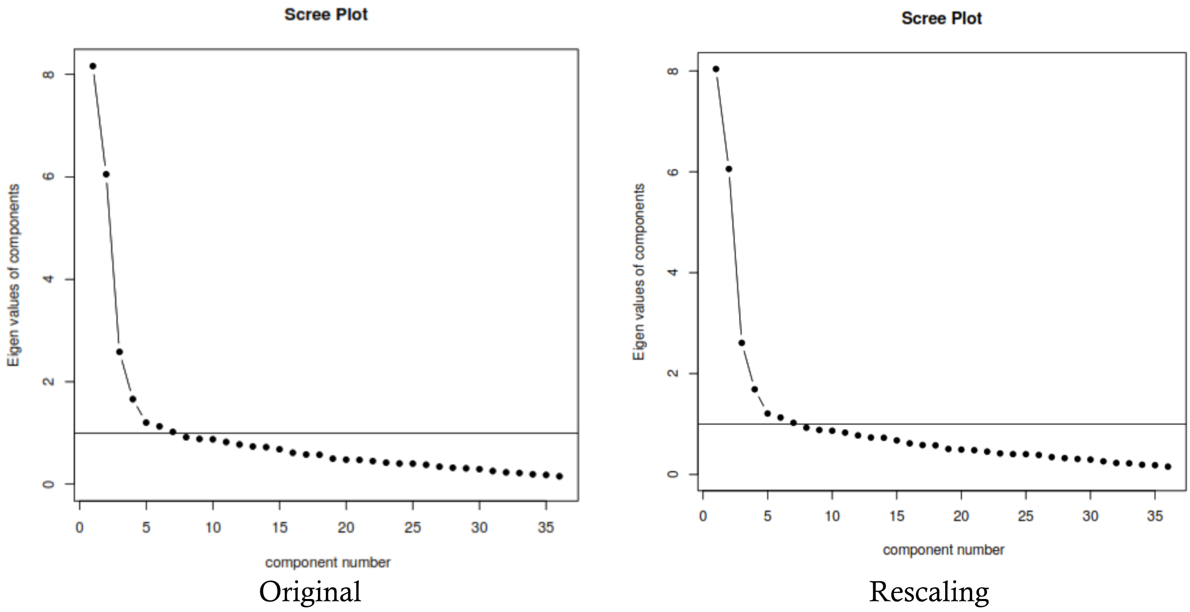
The researchers present the results of the KMO MSA analysis for each Item, raw scores (original), and standardized scores (rescaling) through summated rating scaling in Figure 1 and as follows, in addition to a detailed table for comparison.



Sources: Personal data (2022)

Figure 1. Comparison of KMO MSA for Each Item Original and Rescaling

Based on Figure 1, the KMO MSA graph is obtained based on calculations using the R program in accordance with Table 8, where four items increase, 11 items decrease, and 21 items remain constant. Furthermore, the Eigen Values raw scores (original) and standardized scores (rescaling) through summated rating scaling are presented in Figure 2 as follows.



Sources: Personal data (2022)

Figure 2. Comparison of Eigen Values Original and Rescaling

Screen plots of comparison of Eigen Values raw scores (original) and standardized scores (rescaling) through summated rating scaling explained the connection between many components or formed factors with eigenvalues value in form graph. Figure 2 shows that, at the moment, one component factor formed the curve. It still shows steepness, and at the second and third points of the line curve, it is still sharp. Meanwhile, the 4th, 5th, sixth, and seventh lines curve rather sharply. However, it is different from the 2nd and 3rd line patterns previously. After passing point 7th, the line curve already starts ramps;

increasing to the right will create more ramps. From the explanation, it can be concluded that there are seven components or factors formed in the data before and after being scaled.

Furthermore, based on the results of the analysis using MS Excel and SPSS, the Standard Deviation, Reliability Coefficient, and Standard Error of Measurement (SEm) data were obtained as follows.

Table 9. Results of Standard Deviation, Reliability Coefficient, and Standard Error of Measurement (SEm)

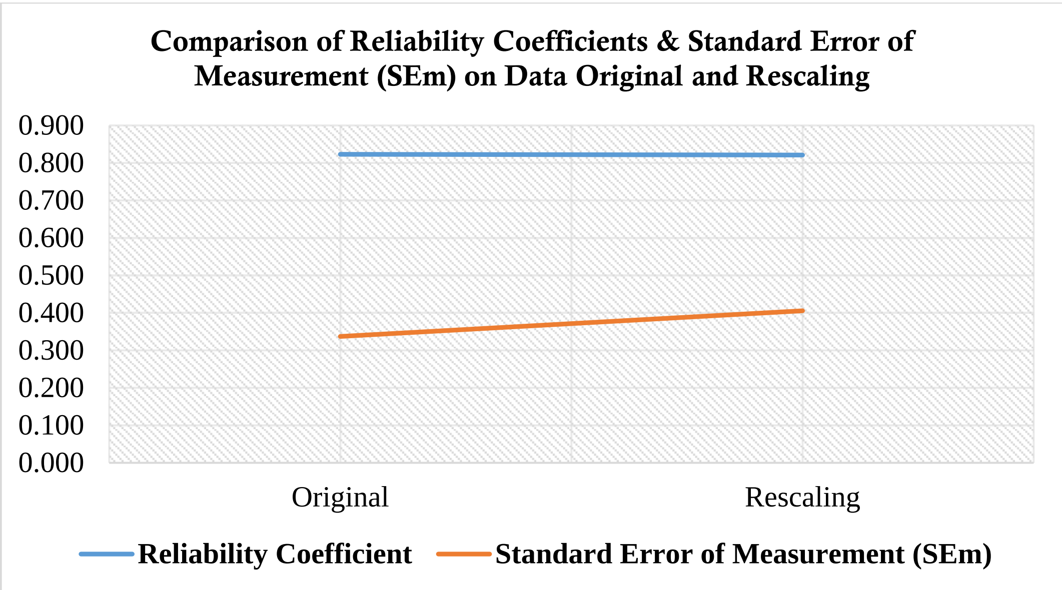
No	Information	Results Original	Results Rescaling
1	Standard Deviation	0.801	0.959
2	Reliability Coefficient	0.823	0.821
3	Standard Error of Measurement (SEm)	0.337	0.406

Sources: Personal data (2022)

Table 9 shows the results of the calculation reliability instrument. Overall, using this formula, we obtained a reliability coefficient of 0.823 with a SEm of 0.337 and a Standard Deviation of 0.801 on the original data. Meanwhile, reliability is 0.821 with a SEm of 0.406 and a Standard Deviation of 0.959 on data that has been transformed into a z score. Reliability of every instrument Likert type raw scores (original) and standardized scores (rescaling) through summated rating scaling process No experience This significant change is in line with research results (Setiawati et al., 2013). Although the reliability of the original data is taller than the existing data scale, the difference is very small and lacks significance. When considering the Standard Error of Measurement (SEm), apparently, SEm on the existing data scale tends to be taller. After reliability on each dimension is obtained, the next step is to count the reliability instrument Likert type using the formula reliability stratified alpha composite (Setiawati et al., 2013).

Chart Comparison Coefficient Reliability and Standard Error of Measurement (SEm)

Below are presented the results of the Comparison Reliability Coefficient and Standard Error of Measurement (SEm) on raw scores (original) and standardized scores (rescaling) through summated rating scaling attitude reason critical Vocational school students are presented in the following picture.



Sources: Personal data (2022)

Figure 3. Comparison of Reliability and Standard Error of Measurement (SEm)

In Table 9 and Figure 3, the Cronbach's Alpha reliability raw scores (original) and standardized scores (rescaling) through summated rating scaling, it can be seen that the empirical test consists of 36 questions with a Cronbach's Alpha value of 0.823 exceeding the minimum limit of 0.60 so the instrument is reliable in the high-reliability category (Steenbergen-Hu et al., 2020; Liu & Cohen, 2021). The questions in the questionnaire relating to measuring critical reasoning attitudes in Muhammadiyah Vocational School students in Yogyakarta are considered reliable and consistent.

The empirical test results show that the critical reasoning instrument (questionnaire) is considered appropriate and standard, with a validity level of 0.91 in the "high" category, reliability of each Item of more than 0.60 in the "strong" category, and factor relationships formed from 36 statement item. This information becomes the basis for researchers in continuing empirical tests.

The calculation results in SEm show that Overall, using this formula, we obtained a SEm of 0.337 on the original data and a SEm of 0.406 on data that had been transformed into a z score. The reliability analysis of the measurement results showed that there was a decrease in the reliability coefficient after the data was transformed into a z-score. Although the decrease in the reliability coefficient was very small, the difference was not clear enough. However, after the measurement results were further analyzed in SEm, there was an increase in SEm in the transformed data. This increase looks quite high, especially for instruments that have high reliability. These results show that scaling does not increase the reliability score but increases the SEm.

Conclusion

Critical reasoning skills play a crucial role in effectively addressing everyday challenges and making informed decisions. These skills enable individuals, particularly students, to derive accurate conclusions grounded in facts, solve problems systematically, and propose effective solutions. The evaluation of students' critical reasoning attitudes is equally important, as it involves measuring non-cognitive dimensions to reflect the qualifications of the attributes being assessed.

Assessment methodologies are diverse, ranging from nominal to ratio scales. Ordinal data derived from Likert-type scales are prevalent in educational assessments but pose limitations for parametric statistical analysis due to their ranking nature. This has led to ongoing debates about transforming ordinal data into interval data to facilitate more robust statistical analyses. One approach to address this is through summated rating scales, known for their reliability and validity in measuring latent constructs.

The present study utilized an adapted instrument with 36 validated and reliable items to assess the critical reasoning attitudes of 204 vocational school students. Data were collected via a questionnaire and analyzed using summated rating scaling to compare the results of KMO MSA, Eigen Value, reliability, and Standard Error of Measurement (SEm) raw scores (original) and standardized scores (rescaling) through summated rating scaling. The analysis involved detailed statistical calculations using tools like Microsoft Excel, SPSS 26, and the R Program.

The findings show that the use of summated rating scaling does not significantly alter the psychometric properties of the instrument. The KMO MSA values for the original and rescaled scores through summated rating scaling were relatively consistent, indicating that the transformation process had minimal impact on the fit of the factor analysis. The rescaled scores provide a more subtle variation in the data, which may improve the interpretation of the results, but does not noticeably increase or decrease the overall reliability and validity of the instrument. This is because the data was not transformed to an interval scale, so the standardized scores retain ordinal characteristics.

In conclusion, although summated rating scaling can refine the granularity of data and potentially affect descriptive statistics and interpretation, its impact on fundamental psychometric properties, such as reliability and validity, appears limited. However, to obtain more accurate results and a truly linear scale, it is recommended that researchers and educators transform the data to an interval scale using the

Rasch model. Thus, the measurement process can be more precise without compromising the integrity of the measurement instrument.

Acknowledgment

We would like to express our appreciation to all those who have played a role in this research and contributed to the writing of this article. This article is a scientific paper prepared by the first author as a student. Our thanks also go to Universitas Negeri Yogyakarta and Universitas Sarjanawiyata Tamansiswa Yogyakarta for their support.

Conflict of Interest

The authors declare that there is no conflict of interest in this research.

Authors Contribution

NDA is fully responsible for the content of the article, starting from the preparation of instruments, data collection, data analysis, to reporting research results. MH played a role in the preparation of research instruments and provided input and improvements to the article manuscript. YP contributed to the preparation of instruments, data collection, and provided suggestions and improvements, especially in the research results section in the article manuscript. AS helped prepare the instrument, collected data, and provided suggestions related to grammar in the manuscript. FAS designed the content concept in the article manuscript, while HR played a role in preparing the instrument and providing suggestions and improvements to the article manuscript.

References

Abduqodirovich, M. A. (2023). "Teaching foreign languages in the context of sustainable development: best practices, problems and opportunities" picularities and importance of critical thinking. *International Scientific and Current Research Conferences*, *1*(1), 355–358. https://doi.org/10.55640/iscrc-3214

Alderson, J. Charles & Wall, D. (1992). Summated Rating Scale Construction: An Introduction. Series: Quantitative Applications in the Social Sciences. In *Sage Publications, Inc.* https://doi.org/10.20595/jjbf.19.0_3

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Brooks/Cole.

Alotaibi, K. (2017). The Relationship Between Self-Regulated Learning and Academic Achievement for a Sample of Community College Students at King Saud University. *Education Journal*, *6*(1), 28. https://doi.org/10.11648/j.edu.20170601.14

Amin, S., Kartowagiran, B., & Inang, P. (2018). A factor analysis of an instrument for measuring physical abuse experience of students at school. *REID (Research and Evaluation in Education)*, *4*(1), 22–34. https://doi.org/10.21831/reid.v4i1.18831

Azwar, Saifuddin. (2014). *Reliabilitas dan validitas*. Pustaka Pelajar.

Brown, T. C., & Daniel, T. C. (1990). *Scaling of ratings: Concepts and methods*. https://www.fs.usda.gov/rm/pubs_rm/rm_rp293.pdf

Carifio, J., & Perla, R. (2008). Resolving the 50-year debate around using and misusing Likert scales. In *Medical Education*, 42(12), 1150–1152. https://doi.org/10.1111/j.1365-2923.2008.03172.x

Christmann, A., & Aelst, S. Van. (2006). *Robust estimation of Cronbach's alpha*. *97*, 1660–1674. https://doi.org/10.1016/j.jmva.2005.05.012

Clark, T., Foster, L., & Bryman, A. (2019). *Analysing quantitative data. How to do your social research project or dissertation*. Oxford University Press. https://doi.org/10.1093/hepl/9780198811060.003.0013

Cromwell, L. S. (1992). *Assessing critical thinking. New Directions for Community Colleges.* Wiley.

Crossley, J. (2016). *Assessing the non-cognitive domains*. Sage Publications.

Crossley, J. (2017). *Assessing the non-cognitive domains: measuring what matters well*. Routledge.

Effendi, H., Soenarto, S., & Sofyan, H. (2015). The effectiveness of web-based interactive blended learning model in electrical engineering courses. *REID (Research and Evaluation in Education)*, *1*(2), 175–185. https://doi.org/10.21831/reid.v1i2.7140

Fajri, F. R., & Amir, M. F. (2022). Math self-regulated learning assisted by metacognitive support by reviewing sex differences in mathematics anxiety. *REID (Research and Evaluation in Education)*, *8*(2), 100–113. https://doi.org/10.21831/reid.v8i2.49157

Faralina, A. A. N., Kadri, A. S., & Yap, A. N. S. (2016). Validity and reliability of pre-class reading tasks for waves and optics (PRT-WO). *REID (Research and Evaluation in Education)*, *2*(1), 42–52. https://doi.org/10.21831/reid.v2i1.8466

Flinton, D., & Malamateniou, C. (2020). *Quantitative methods and analysis*. Pearson.

Golden, B. (2023). Enabling critical thinking development in higher education through the use of a structured planning tool. *Irish Educational Studies*, *42*(4), 949–969. https://doi.org/10.1080/03323315.2023.2258497

Hajaroh, M., Rukiyati, Purwastuti, L. A., & Nurhayati, R. (2021). Development of the evaluation instrument of the child-friendly school policy in elementary schools. *International Journal of Instruction*, *14*(3), 327–340. https://doi.org/10.29333/iji.2021.14319a

Hamdi, S., Suganda, I. A., & Hayati, N. (2018). Developing higher-order thinking skill (HOTS) test instrument using Lombok local cultures as contexts for junior secondary school mathematics. *REID (Research and Evaluation in Education)*, *4*(2), 126–135. https://doi.org/10.21831/reid.v4i2.22089

Hasanatin, M., & Rohaeti, E. (2021). How are the characteristics of integrated assessment instruments to measure critical thinking skills and scientific attitudes of high school students? *Jurnal Penelitian Pendidikan IPA*, *7*(SpecialIssue), 359–364. https://doi.org/10.29303/jppipa.v7ispecialissue.1082

Hudha, S. A., & Mardapi, D. (2018). Developing an instrument for measuring the spiritual attitude of high school students. *REID (Research and Evaluation in Education)*, *4*(1), 35–44. https://doi.org/10.21831/reid.v4i1.20304

Jamieson, S. (2004). Likert scales: How to (ab)use them. In *Medical Education* (Vol. 38, Issue 12, pp. 1217–1218). https://doi.org/10.1111/j.1365-2929.2004.02012.x

Jian-hua, C. (2005). *The application of SPSS (software) in the sports measurement & survey.* Sports Sciences Researches.

Kalkbrenner, M. T. (2021). A practical guide to instrument development and score validation in the social sciences: The measure approach. *Practical Assessment, Research and Evaluation*, *26*(1), 1–18. https://doi.org/https://doi.org/10.7275/svg4-e671

Kampen, J. K. (2019). Reflections on and test of the metrological properties of summated rating, Likert, and other scales based on sums of ordinal variables. *Measurement: Journal of the International Measurement Confederation*, *137*, 428–434. https://doi.org/10.1016/j.measurement.2019.01.083

Kim, M., & Pegg, J. (2019). Case analysis of children's reasoning in problem-solving process. *International Journal of Science Education*, *41*(6), 739–758. https://doi.org/10.1080/09500693.2019.1579391

Kumolohadi, R., Mangunsong, F., & Suleeman, J. (2021). Development and validation of ethical awareness scale among university students. *Jurnal Pengukuran Psikologi dan Pendidikan Indonesia*, *10*(1), 20–31. https://doi.org/10.15408/jp3i.v10i1.17216

Kurniawan, N. I. A., & Munadi, S. (2019). Analysis of the quality of test instrument and students' accounting learning competencies at vocational school. *Jurnal Penelitian dan Evaluasi Pendidikan*, *23*(1), 68–75. https://doi.org/10.21831/pep.v23i1.22484

Liu, J., & Cohen, J. (2021). Measuring teaching practices at scale: A novel application of text-as-data methods. *Educational Evaluation and Policy Analysis*, *43*(4), 587–614. https://doi.org/10.3102/01623737211009267

Lorenzo-Seva, U., & Ferrando, P. J. (2021). MSA: The Forgotten Index for Identifying Inappropriate Items Before Computing Exploratory Item Factor Analysis. *Methodology*, *17*(4), 296–306. https://doi.org/10.5964/meth.7185

MacInnes, J. (2017). *An Introduction to Secondary Data Analysis with IBM SPSS Statistics*. Sage Publications.

Meyer, E. P., Kaiser, H. F., Cerny, B. A., & Green, B. F. (1977). MSA for a special spearman matrix. *Psychometrika*, *42* (1), 153-156. https://doi.org./10.1007/BF02293753

Nkansah, B. K. (2018). On the Kaiser-Meier-Olkin's Measure of Sampling Adequacy. *Mathematical Theory and Modeling*, *8(7)*, 52–76. https://iiste.org/Journals/index.php/MTM/article/view/44386/45790

Nolan, M. M., Beran, T., & Hecker, K. G. (2012). Surveys assessing students' attitudes toward statistics: A systematic review of validity and reliability. *Statistics Education Research Journal*, *11*(2), 103–123. https://doi.org/10.52041/serj.v11i2.333

Okoye, K., Rodriguez-Tort, J. A., Escamilla, J., & Hosseini, S. (2021). Technology-mediated teaching and learning process: A conceptual study of educators' response amidst the Covid-19 pandemic. *Education and Information Technologies*, *26*(6), 7225–7257. https://doi.org/10.1007/s10639-021-10527-x

Prasasti, I. H., & Istiyono, E. (2018). Developing an instrument of national examination of equivalency education Package C of mathematics subject. *REID (Research and Evaluation in Education)*, *4*(1), 58–69. https://doi.org/10.21831/reid.v4i1.15556

Proches, S. (2016). *Descriptive statistics in research and teaching: are we losing the middle ground? Quality & Quantity*. Springer Netherlands. https://doi.org/10.1007/s11135-015-0256-3

Purnamasari, N., Siswanto, S., & Malik, S. (2020). E-module as an emergency-innovated learning source during the Covid-19 outbreak. *Psychology, Evaluation, and Technology in Educational Research*, *3*(1), 1–8. https://doi.org/10.33292/petier.v3i1.53

Puspitasari, E. D., Susilo, M. J., & Febrianti, N. (2019). Developing psychomotor evaluation instrument of biochemistry practicum for university students of biology education. *REID (Research and Evaluation in Education)*, *5*(1), 1–9. https://doi.org/10.21831/reid.v5i1.22126

Pramudita, K., Rosnawati, R., & Mam, S. (2019). Methods used by mathematics teachers in developing parallel multiple-choice test items in school. *REID (Research and Evaluation in Education)*, *5*(1), 10–20. https://doi.org/10.21831/reid.v5i1.22219

Rachmantika, A. R., & Wardono. (2019). The role of students' critical thinking skills in learning mathematics with problem-solving. *Prosiding Seminar Nasional Matematika*, *2*(1), 441. https://journal.unnes.ac.id/sju/index.php/prisma/article/view/29029

Rasool, S., Cerchione, R., Salo, J., Ferraris, A., & Abbate, S. (2021). Measurement of consumer awareness of food waste: construct development with a confirmatory factor analysis. *British Food Journal*, *123*(13), 337–361. https://doi.org/10.1108/BFJ-02-2021-0160

Retnawati, H. (2016). *Analisis kuantitatif instrumen penelitian (panduan peneliti, mahasiswa, dan psikometrian)*. Parama Publishing.

Rosana, D., Widodo, E., Setianingsih, W., & Setyawarno, D. (2020). Developing Assessment Instruments of PISA Model to Measure Students' Problem-Solving Skills and Scientific Literacy in Junior High Schools. *Jurnal Pendidikan Sains Indonesia*, *8*(2), 292–305. https://doi.org/10.24815/jpsi.v8i2.17468

Sabon, Y. O. S., Istiyono, E., & Widihastuti, W. (2022). Developing "Pancasila Student Profile" instrument for self-assessment. *Jurnal Penelitian dan Evaluasi Pendidikan*, *26*(1), 37–46. https://doi.org/10.21831/pep.v26i1.45144

Santoso, P. H., Istiyono, E., & Haryanto. (2022). Principal Component Analysis and Exploratory Factor Analysis of the Mechanical Waves Conceptual Survey. *Jurnal Pengukuran Psikologi Dan Pendidikan Indonesia*, *11*(2), 208–224. https://doi.org/10.15408/jp3i.v11i2.27488

Setiawan, R., Wagiran, W., & Alsamiri, Y. (2024). Construction of an instrument for evaluating the teaching process in higher education: Content and construct validity. *REID (Research and Evaluation in Education)*, *10*(1), 50–63. https://doi.org/10.21831/reid.v10i1.63483

Setiawati, F. A. (2013). Penskalaan tipe likert dan thurstone dengan teori klasik dan modern: stidu pada instrumen multiple intelligences. *Yogyakarta, Universitas Negeri Penelitian, Lembaga Pengabdian, dan Masyarakat, Kepada Masyarakat 2013,* 42.

Setiawati, F. A. (2014). Perbandingan berbagai metode penskalaan yang dikembangkan Thurstone. *Jurnal Penelitian Ilmu Pendidikan*, *7*(1), 86–93. https://doi.org/10.21831/jpipfip.v7i1.3114

Setiawati, F. A., Mardapi, D., & Azwar, S. (2013). Penskalaan Teori Klasik Instrumen Multiple Intelligences Tipe Thurstone Dan Likert. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *17*(2), 259–274. https://doi.org/10.21831/pep.v17i2.1699

Sharma, B. (2016). A focus on reliability in developmental research through Cronbach's Alpha among medical, dental and paramedical professionals. *Asian Pacific Journal of Health Sciences*, *3*(4), 271–278. https://doi.org/10.21276/apjhs.2016.3.4.43

Shevlin, M., Miles, J. N. V., & Bunting, B. P. (1997). Summated rating scales: a monte carlo investigation of the effects of reliability and collinearity in regression models. *Elsevier Science*, *23*(4), 665–676. https://doi.org/10.1016/S0191-8869(97)00088-3

Spector, P. E. (1976). Choosing Response Categories for Summated Rating Scales. *Journal of Applied Psychology*, *61*(3), 374–375. https://doi.org/10.1037/0021-9010.61.3.374

Spector, P. E. (1992). *Summated rating scale construction: an introduction*. Sage Publications.

Steenbergen-Hu, S., Olszewski-Kubilius, P., & Calvert, E. (2020). The Effectiveness of Current Interventions to Reverse the Underachievement of Gifted Students: Findings of a Meta-Analysis and Systematic Review. *Gifted Child Quarterly*, *64*(2), 132–165. https://doi.org/10.1177/0016986220908601

Stępień, K., Janecki, D., & Adamczak, S. (2021). Statistical study of the method of radius-change sphericity measurements. *Sensors*, *21*(13), 1–19. https://doi.org/10.3390/s21134563

Sugiyono, S., Mardapi, D., & Suryadarma, I. G. P. (2017). Developing assessment instruments for the sensory acceptability of food products. *REID (Research and Evaluation in Education)*, *3*(1), 92–105. https://doi.org/10.21831/reid.v3i1.13940

Sumin, Sukmawati, F., Setiawati, F. A., & Asmawi, S. (2022). The Impact of Z-Score Transformation Scaling on the Validity, Reliability, and Measurement Error of Instrument SATS-36. *Jurnal Pengukuran Psikologi Dan Pendidikan Indonesia*, *11*(2), 166–180. https://doi.org/10.15408/jp3i.v11i2.26591

Suryabrata, S. (2002). *Pengembangan alat ukur psikologis*. Andi Offset.

Syahputra, W., Yundianto, D., Baydhowi, & Paramita, N. D. (2023). The Indonesian Democratic Citizenship Attitude Scale: Translation, Cross-Cultural Adaptation, and Psychometric Properties. *Jurnal Pengukuran Psikologi Dan Pendidikan Indonesia*, *12*(2), 191–205. https://doi.org/10.15408/jp3i.v12i2.28414

Taber, K. S. (2018). The Use of Cronbach' Alpha When Developing and Reporting Research Instruments in Science Education. *Research in Science Education*, *48*(6), 1273–1296. https://doi.org/10.1007/s11165-016-9602-2

Teng, Y., & Wang, X. (2021). The effect of two educational technology tools on student engagement in Chinese EFL courses. *International Journal of Educational Technology in Higher Education*, *18*(1), 0–15. https://doi.org/10.1186/s41239-021-00263-0

Tyler, F., & Michael, W. B. (1958). An Empirical Study of the Comparability of Factor Structure when Unities and Communality Estimates are Used. *Educational and Psychological Measurement*, *18*(2), 347–354. https://doi.org/10.1177/001316445801800021

Vijayendra, K., & Fantone, A. (2023). Recent Trends of Quantitative Approaches in Different Sectors: A Concise Review. *Asian Journal of Education and Social Studies*, *41*(2), 22–34. https://doi.org/10.9734/ajess/2023/v41i2891

Wardani, R. E. A., Supriyoko, S., & Prihatni, Y. (2018). Developing assessment model for bandel attitudes based on the teachings of Ki Hadjar Dewantara. *REID (Research and Evaluation in Education)*, *4*(2), 117–125. https://doi.org/10.21831/reid.v4i2.21942

Wiyono, M., Solihin, F., & Putro, S. S. (2017). Aplikasi Penilaian Kuliah Kerja Nyata Universitas Trunojoyo Madura Menggunakan Metode Rating Scale. *Jurnal Imliah Rekayasa*, *10*(1), 23–33. https://doi.org/10.21107/rekayasa.v10i1.3601

Appendix

1. Instrument (36 Item): http://gg.gg/Instrument_36_Item
2. Calculation Summated Rating Scaling (36 Item): https://bit.ly/36_Item_Summated_Rating
3. Data SD, SEm, and Reliability Analysis: http://gg.gg/3-SD_SEm_and_Reliability_Analysis
4. SPSS and R Program Analysis Results: http://gg.gg/SPSS_and_R_Program_Analysis_Results